6

THE BAYESIAN APPROACH TO STATISTICS

ANTHONY O'HAGAN

INTRODUCTION

By far the most widely taught and used statistical methods in practice are those of the frequentist school. The ideas of frequentist inference, as set out in Chapter 5 of this book, rest on the frequency definition of probability (Chapter 2), and were developed in the first half of the 20th century. This chapter concerns a radically different approach to statistics, the Bayesian approach, which depends instead on the subjective definition of probability (Chapter 3). In some respects, Bayesian methods are older than frequentist ones, having been the basis of very early statistical reasoning as far back as the 18th century. Bayesian statistics as it is now understood, however, dates back to the 1950s, with subsequent development in the second half of the 20th century. Over that time, the Bayesian approach has steadily gained ground, and is now recognized as a legitimate alternative to the frequentist approach.

This chapter is organized into three sections. The first gives an outline of the Bayesian method. The second section contrasts the Bayesian and frequentist schools, linking their differences to fundamental differences over the interpretation of probability, and argues that the Bayesian approach is more consistent and reflects better the true nature of scientific reasoning. The final section addresses various features of modern Bayesian methods that provide some explanation for the rapid increase in their adoption since the 1980s.

BAYESIAN INFERENCE

We first present the basic procedures of Bayesian inference.

Bayes's Theorem and the Nature of Learning

Bayesian inference is a process of learning from data. To give substance to this statement, we need to identify who is doing the learning and what they are learning about.

Terms and Notation

The person doing the learning is an individual scientist, analyst, or decision maker who wishes to learn from the data. Where we need to refer to this person explicitly, we will call him or her "You." The choice of word emphasizes the fact that Bayesian inference is concerned with the knowledge of a particular person, and so is intrinsically subjective, but the capital letter "Y"

when discussing general principles distinguishes this person from "you, the reader" and shows that we are referring to an abstract or arbitrary person.

As with other approaches to statistics, the object of analyzing the data is to make inferences about some unknown parameters. It is conventional to denote parameters by Greek letters, and when discussing general principles we denote them by θ. In context, θ may represent a single parameter or more generally a collection of parameters. The data are usually denoted by Roman letters, and in general discussion we use the symbol x. In Bayesian statistics, You use the data x to learn about the parameters θ. Your beliefs and knowledge about θ are *updated* in this learning process.

We, therefore, need notation and terminology to describe Your state of knowledge before and after learning from the data. We refer to knowledge before observing the data as *prior* information, and to that obtaining after observing the data as *posterior* information. The words "prior" and "posterior" are relative to the data under consideration.

The description of prior or posterior knowledge in the Bayesian framework is a probability distribution. Your prior distribution for θ is denoted by $f(\theta)$ and Your posterior distribution by $f(\theta\,|\,x)$. For the purposes of this chapter, we will take these to be probability density functions, since in the great majority of applications the parameters are continuous variables.[1] Thus, the state of knowledge after observing the data is distinguished from that before observing the data simply by conditioning on x.[2]

The prior distribution is a complete description of Your prior knowledge, in the sense we can derive from it Your prior probability that θ lies in any set of interest. Similarly, $f(\theta\,|\,x)$ is a complete description of Your posterior information about θ. This conceptual description of the prior distribution will suffice as we explore the mechanics of Bayesian inference. We will consider the prior information in more detail in the section Prior Distributions.

Bayes's Theorem

Bayes's theorem (written "Bayes' theorem" by some and named after the 18th-century clergyman and mathematician Thomas Bayes) is the formula for deriving the posterior distribution. In the form used in Bayesian statistics, the theorem can be simply expressed as

$$f(\theta\,|\,x) \propto f(\theta)\, f(x\,|\,\theta). \tag{6.1}$$

To understand this formula, first note that on the left-hand side of (6.1) is the posterior density $f(\theta\,|\,x)$, whereas on the right-hand side is the product of two terms, one of which is the prior density $f(\theta)$. The other term, $f(x\,|\,\theta)$, is the probability distribution for the data, conditional on the parameter θ. This distribution also appears in all other approaches to statistics, and in particular in frequentist theory. When thought of as a function of the unknown parameters θ (and for fixed data x), it is called the *likelihood function* (or simply the likelihood); see Chapter 5.

Note next that the left- and right-hand sides of (6.1) are linked by the proportionality symbol, "$\propto$." The theorem therefore says that the posterior density is *proportional to* the product of the prior density and the likelihood. We need a proportionality symbol here rather than an equals sign because the posterior density, like any density function, must have its integral (i.e., the area under the curve) equal to 1. If we simply multiply the prior density and the likelihood function, then the result will not integrate to 1 (except by some remote accident). Therefore, to obtain the posterior density function we must scale the right-hand side by multiplying it by a suitable constant to make it integrate to 1 over the full range of possible values of θ. This is the meaning of proportionality: The posterior density is the product of the prior density and likelihood *except* for a constant (in the sense of not depending on θ) scaling factor.

[1] It is straightforward to adapt everything to deal with discrete-valued parameters, but to do so here in any rigorous way would make the notation and discussion unnecessarily complex.

[2] Strictly, we should explicitly show Your prior information as, say, I. Then, posterior information comprises both I and the observed data x. In this more explicit notation, the prior distribution would be $f(\theta\,|\,I)$ and the posterior distribution $f(\theta\,|\,I,x)$. However, it is usual to suppress the prior information to simplify the notation.

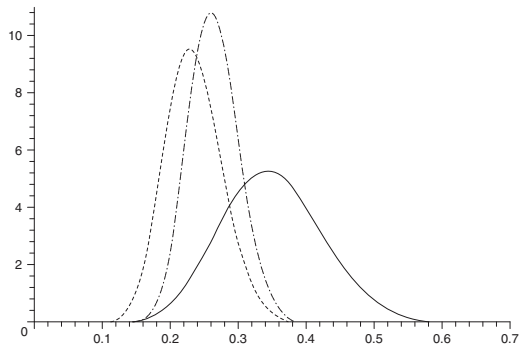

Figure 6.1 Example of mobile phone usage.

NOTE: Solid line: prior density; dashed line: likelihood; dot-dashed line: posterior density.

Learning

The way in which Bayes's theorem operates is best seen through examples. Suppose that You are interested in the proportion of people in the United Kingdom who have used a mobile phone while driving in the last year. If we denote this proportion by θ, then it can take any value in the range 0 to 1.[3] Suppose that we obtain data from a survey of 100 people in the United Kingdom, of whom 23 report having used a mobile phone while driving last year. Figure 6.1 illustrates the use of Bayes's theorem in this case. The solid line is Your prior distribution, which for this example indicates a belief that θ would most probably be in the range 0.3 to 0.4 and is unlikely to lie outside the range 0.2 to 0.5. The dashed line is the likelihood, with the data indicating θ is around the observed frequency of 0.23. The posterior distribution is the dot-dashed line.

Bayes's theorem multiplies the prior density and likelihood. Where either of these is very near zero, the product is near zero, so the posterior density is negligible for $\theta < 0.15$ (because the prior is negligible there) or $\theta > 0.4$ (because the likelihood is negligible there). It covers a narrower range, and so is more informative, than either the prior or the likelihood. The posterior reaches its maximum at $\theta = 0.264$, which represents a compromise between the prior density's maximum at $\theta = 0.35$ and the likelihood's maximum at $\theta = 0.23$. Both the data and prior information have a role in Bayesian inference, and the posterior distribution synthesizes the two sources of information. In this case, the data are more informative than the prior distribution, so this compromise yields a value closer to the data estimate than the prior maximum. We see this also in the fact that the posterior is similar to the likelihood, although the prior information has had some influence in moving the posterior toward larger θ values than the data alone suggest.

This example illustrates very typical behavior of Bayes's theorem.

- The posterior distribution combines the information in both the prior distribution and the likelihood. This typically results in the posterior representing stronger information, and supporting a narrower range of possible values for θ, than either of the separate sources of information.
- The posterior distribution centers around a value that is typically a compromise between the values that are well supported by the prior and by the data separately.
- This compromise also reflects the relative strengths of the prior information and data. The posterior is generally more similar to, and is centered nearer to the center of, the stronger information source.

Figure 6.1 is an example of a triplot, in which the prior, likelihood, and posterior are plotted together on a single graph. When there is just a single unknown parameter, it is a powerful way to see the operation of Bayes's theorem. In practice, however, statistical models nearly always have many more parameters. Bayes's theorem still operates in the same way, but it is no longer so simple to visualize graphically.

Bayes's theorem is the fundamental paradigm for learning from experience, allowing You to update Your prior information to Your posterior information via the evidence in the data. Psychologists have studied how people actually process information, and although we typically do not do so as efficiently as Bayes's theorem dictates, and

[3] Strictly, if N is the population of the United Kingdom, it takes values $0, \frac{1}{N}, \frac{2}{N}, \ldots$, but N is large!

are inclined to make some predictable kinds of judgmental errors, it is clear that people do learn from evidence in broadly this way.

Sequential Learning

We never stop learning. Learning is an ongoing process, and Bayes's theorem reflects this fact in a nice way. Remember that the words "prior" and "posterior" are relative to the data being assimilated. We can apply Bayes's theorem sequentially, to assimilate data piece by piece (or in chunks, as we wish), but then we have to recognize that at any point in this process Your prior distribution should represent the information that is available prior to the particular piece of data that You are about to observe. This, of course, is what You would have called Your posterior distribution after observing the previous piece of data. There is a nice phrase that sums up this analysis: "Today's posterior is tomorrow's prior."[4]

Bayes Estimates and Other Inferences

The basic principle of Bayesian inference is that all inferences are derived from Your posterior distribution. The posterior density $f(\theta\,|\,x)$ expresses all the information that You have about θ after observing the data, and can itself be considered an inference, in the sense of being Your answer to the question, "What do we now know about θ?" For a single parameter, simply drawing its posterior distribution, as in Figure 6.1, provides a clear visualization of that knowledge. It is usual, however, to require more quantitative inferences, such as point estimates, interval estimates, or tests of hypotheses. All of these are derived, in the Bayesian framework, from the posterior distribution.

A point estimate is a single value for θ that represents in some sense a "best guess" in the light of the data and Your prior information. There are several possible choices, depending on what kind of "best" value is required. The posterior median is one kind of Bayesian estimate. We can think of it as a central value, such that the probability that θ is higher than this value equals the probability that it is lower. Another choice is the posterior mean, which is the expected value of θ. This is widely used, and is usually understood when we talk about "Bayes estimates." Finally, the posterior mode is also commonly used, representing the most probable value.[5]

A Bayesian interval estimate is simply an interval having a specified posterior probability. For instance, a 90% interval is a range of values, say $[a,b]$, such that $P(a \leq \theta \leq b\,|\,x) = 0.9$. The usual term for such an interval is a *credible interval*.[6] Bayesian hypothesis testing is particularly straightforward: If You wish to decide, for example, whether to accept the hypothesis that θ is positive, You simply evaluate Your posterior probability that it is true, $P(\theta > 0\,|\,x)$.[7]

For example, in the case of the posterior distribution in Figure 6.1, the median, mean, and mode are, respectively, 0.263, 0.264, and 0.261, so the differences between them are very small and it would not matter in practice which we used. This is because the posterior density is nearly symmetric. In a skewed distribution, the differences between these point estimates will be larger. Next, suppose that You require a 75% credible interval

[4]It can be shown mathematically that Bayesian sequential updating is consistent, in the sense that You will obtain the same posterior distribution by assimilating all the data in one application of Bayes's theorem as You would obtain from applying it sequentially with the same data broken into individual items or blocks of data.

[5]Informally, it might also be called the "most likely" value, but this invites confusion with the frequentist "maximum likelihood estimator," which is quite different. Strictly, for a continuous parameter θ, there is no most probable value since the probability that θ takes any value precisely is zero. However, the mode maximizes the probability that θ will be within a small neighborhood of the estimate.

[6]Although this is a similar term to the frequentist "confidence interval," it is quite different and has a different interpretation; see section "Implications for Inference."

[7]Bayesian methods thereby separate the evaluation of how probable a hypothesis is from any decision whether to "accept" or "reject" it. The probability is a scientific judgment, but to make an accept/reject decision You should take into account the consequences of incorrect decisions. In some situations, You might be willing to accept a hypothesis if its probability is larger than 0.5, but in other situations You may require a much larger probability. For instance, in British criminal law, the accused is judged guilty only if the hypothesis of guilt is proved "beyond all reasonable doubt," whereas in civil law a judgment between two people is made "on the balance of probabilities."

for θ. There are many that we could use, ranging from $[0, 0.289]$ (which takes the 75% lowest possible values) to $[0.239, 1]$ (which uses the 75% highest values). The shortest possible 75% credible interval is, however, $[0.220, 0.305]$.[8] Finally, if You wish to decide whether to accept the hypothesis that $\theta < 0.25$, the posterior probability is $P(\theta < 0.25 \,|\, x) = 0.36$. So it is more probable, according to Your posterior distribution, that $\theta > 0.25$, but there is still substantial uncertainty about this hypothesis.[9]

Prior Distributions

The prior distribution is an intrinsic part of the Bayesian approach and the most obvious feature that distinguishes it from the frequentist approach. Much of the controversy about which inference paradigm is better has centered on the prior distribution. We will discuss the main arguments in this debate in the section Parameters as Random Variables, but first we consider how the prior distribution is specified in practice.

Elicitation

The most basic way to specify Your prior distribution for θ is a process known as *elicitation*. The word derives from the fact that elicitation usually involves an external facilitator who constructs the prior distribution to represent the knowledge of the person whose prior information is to be elicited. In practice, whereas the analysis of the data, construction of the posterior distribution, and derivation of appropriate inferences might be carried out by a statistician, the person whose prior information is to be elicited (that we have called You) will typically not be knowledgeable about statistics. We therefore consider elicitation to be a dialogue between the facilitator (someone with expertise in statistics and the elicitation of expert knowledge) and the subject-matter expert (You).[10]

In response to questions from the facilitator, You will specify particular features of Your prior knowledge. For instance, You might specify Your prior median and a prior 80% credible interval. The facilitator then constructs a prior distribution to represent Your stated beliefs. The skill of the facilitator lies in deciding which features of Your prior knowledge to ask about and how to ask those questions without biasing the answers. It is important to be aware of the considerable research in psychology concerning the ways that people respond to questions about uncertainty (see, for instance, O'Hagan et al., 2006).

Elicitation is not a precise process. First, it is difficult for You to think quantitatively about Your prior knowledge, and we cannot expect Your answers to be precise. For instance, if You are asked to specify Your prior probability that $\theta < 0.25$ in the mobile phone example, You might feel that θ is probably larger than 0.25, but how probable? You might say $P(\theta < 0.25) = 0.1$, but if pressed by the facilitator You might be perfectly happy with any value between 0.07 and 0.15 for this probability. The second source of imprecision is that You can only specify a relatively small number of features of Your prior distribution, partly because time is always limited and partly because the task becomes more complex as more questions are asked. In choosing a distribution to represent Your stated beliefs, the facilitator is making an arbitrary choice, and in reality there is a whole range of prior distributions that might fit Your statements equally well.

Because elicitation is imprecise, there is imprecision in the posterior distribution, and in inferences derived from it. It is important in practice to explore how robust any derived inferences might be to perturbing the prior distribution.

Fortunately, this prior imprecision often does not matter, because the posterior distribution is

[8]The shortest credible interval for any given probability of containing the true values is known as the *highest density interval*, because it is found by including in the interval all those values of θ having highest (posterior) density.

[9]However, Your prior probability $P(\theta < 0.25) = 0.09$ has been greatly increased by the data.

[10]The separation of roles is not always necessary, and certainly You could elicit Your prior distribution by playing both parts in the dialogue. Nevertheless, in situations where prior information is substantial and the problem of sufficient importance, the use of an experienced facilitator is advisable. In some projects, the opinions of several experts might be elicited, either individually or in a group.

almost unaffected by varying the prior distribution over quite a wide range. This is the case when the data are sufficiently strong. We have seen that Bayes's theorem synthesises the two sources of information by giving more weight to the stronger source. When the data are far more informative than the prior distribution,[11] the posterior distribution is almost entirely determined by the likelihood, and varying the prior distribution produces little effect. This is, therefore, when the posterior inferences will be robust to imprecision in the prior distribution.

Noninformative Priors

Another way to look at this is to say that we can avoid the posterior distribution being dependent on the prior if we make the prior information very weak. Prior distributions that express negligible prior knowledge have been given a huge variety of names, but we will use here the term "noninformative."[12] Several justifications have been proposed for considering such prior distributions.

1. Those who like the elegance of the Bayesian approach (with particular reference to its benefits over frequentist methods as discussed in the section Contrast With Frequentist Inference), yet are concerned about criticisms of the use of prior information, see noninformative prior distributions as a way to achieve the Bayesian benefits without the Bayesian prior.

2. To study the relationship between Bayesian and frequentist methods, it can be useful to formulate noninformative prior distributions, since this should bring the two approaches as close together as possible.

3. When prior information is genuinely weak compared with the data, so that the posterior distribution should not anyway be sensitive to the particular distribution that we use, then a noninformative prior is a convenient choice that avoids the need to go through a process of elicitation.

There has been a substantial amount of research into defining noninformative priors for various kinds of parameters in various models, but this is a contentious topic (see Berger, 2006; Goldstein, 2006; and discussions following those articles). There is no consensus over which of the competing recipes best represent prior "ignorance" in any given situation, and indeed many scholars would argue that complete ignorance never exists and there is no such thing as a totally noninformative prior. Nevertheless, in most cases the various noninformative priors that have been proposed in the literature for any given problem should lead to essentially the same posterior inferences, provided the data are not themselves weak.

The author's view is that genuine, informative prior distributions should be used wherever substantive prior information exists, but that when prior information is truly weak relative to the data then so-called noninformative prior distributions play a useful role in Bayesian inference (following the third justification above). In problems with several unknown parameters, it is rare for there to be useful prior information about all the parameters, so it is sensible to make efforts to formulate proper prior distributions for those parameters where genuine prior information exists, and to place conventional noninformative prior distributions on the others.

Data-Based Priors

Prior information often includes other data, say y, separate from the particular data x being analyzed. Then, in principle, we can say that "today's" prior distribution (before observing x) is "yesterday's" posterior distribution (after observing y). It might then be written $f(\theta \mid y)$ and could be derived using Bayes's theorem from "yesterday's" likelihood $f(y \mid \theta)$ and "yesterday's" prior $f(\theta)$.

In practice, though, it is not simple to deal with "prior data" in this way. First, the problem of specifying "today's" prior distribution has

[11] We are using the term *informative* here in the sense of the discussion of the triplot (Figure 6.1). We refer here specifically to the situation where the prior distribution is very much broader and flatter than the likelihood.

[12] Some of the other names are "weak," "reference," "default," "vague," "objective," or "ignorance" priors.

simply been deferred to that of specifying "yesterday's" prior, and it will generally be difficult to think about what prior distribution would have applied in the hypothetical state of knowledge prior to observing y. Second, y often relates only indirectly to θ. This is the case when prior information relates to knowledge of *similar* problems. For instance, if required to assess a prior distribution for the efficacy θ of some new drug, You may have prior experience of the performance of similar drugs. To use such data formally alongside x, it is necessary first to formulate its relationship to θ in the form of "yesterday's" likelihood $f(y|\theta)$ (e.g., by explicitly formulating some assessment of similarity between the new and old drugs). Such complications mean that it is often no easier to make explicit use of prior data than to elicit Your current prior distribution (so incorporating y implicitly).

Contrast With Frequentist Inference

Where appreciable prior information exists, perhaps the most significant difference between Bayesian and frequentist methods is the ability of the Bayesian analysis to make use of that additional information in the form of the prior distribution. As a result, Bayesian methods will typically produce stronger inferences from the same data. Furthermore, the prior information allows the Bayesian analysis to be more responsive to the context of the data. However, the prior distribution is also the focus of opposition to Bayesian methods from adherents of the frequentist philosophy. Frequentists regard its use as unscientific, so do not believe that such stronger or more responsive inferences can be obtained legitimately.

Parameters as Random Variables

Although the use of a prior distribution does distinguish Bayesian methods from frequentist methods, we have seen that some users of Bayesian ideas attempt to nullify the prior information by using noninformative priors. Even where genuine prior distributions are employed, they may have very little impact on the inferences. A better defining characteristic for the Bayesian approach is the willingness to treat unknown parameters as random variables.

The Nature of Probability and Uncertainty

We can only have a posterior distribution if θ is considered as a random variable. In frequentist statistics, parameters cannot be random variables, and it is not legitimate to make probability statements about them. This, more than whether one feels discomfort with the use of prior information, is what makes frequentist inference fundamentally different from Bayesian inference.[13]

Underlying this distinction is a still more fundamental difference over what probability means. Frequentist inference is so called because it relies on the frequency interpretation of probability, so that every probability is defined as the long run relative frequency with which events of that type occur under repeated observation. Probability statements cannot be made about parameters because they cannot meaningfully be considered as repeatable. In any statistical problem, we have data that are generally sampled from some population or data-generating process that is repeatable. We can consider drawing samples indefinitely from such a process, and so x is a random variable within the frequency formulation, and its distribution $f(x|\theta)$ is well defined in terms of frequency probabilities.[14] However, θ represents the unknown features of that data-generating process. They are fixed and specific to this problem. θ is unique and cannot be considered part of a repeatable sequence, so we cannot meaningfully assign frequency probabilities to it.[15]

[13]In some problems, frequentist statistics makes use of "random effects" formulations, in which some parameters in an analogous "fixed effects" model become random variables. However, the random effects are then not then treated as parameters, and inference cannot be made about individual random effects.

[14]Technically, in frequentist inference, because θ is not a random variable we do not formally *condition* on its value, and hence the notation $f(x|\theta)$ is strictly incorrect. It is usual to write it instead as $f(x;\theta)$ or $f_\theta(x)$.

[15]Even if we could conceive of a collection of data-generating processes, the one under study is not randomly sampled from that collection and inevitably has its own characteristics that make it not comparable with the others.

Philosophically, different kinds of uncertainty are associated with x and θ. The data are subject to random variability, and the associated uncertainty is termed *aleatory* (from the Latin "alea" for a die). Parameters are not random (in the everyday sense of this word), but they are uncertain. The uncertainty in this case arises from a lack of knowledge and is termed *epistemic* (from the Greek "episteme" for science or knowledge).[16] Frequency probability is only applicable to quantify aleatory uncertainties. In contrast, the subjective or personal interpretation of probability defines Your probability for an event as a measure of Your degree of belief in the assertion that the event will occur. This definition clearly applies to any uncertain event, whether the uncertainty is epistemic or aleatory.[17]

The willingness to express uncertainty about θ through probabilities, and to assign a probability distribution to θ either before or after observing x, means that Bayesian inference is intrinsically based on the subjective formulation of probability.

Implications for Inference

Frequentist inference cannot make probability statements about parameters, yet it often appears to do just that.

Consider a hypothesis testing problem, where the inference question is to decide whether to accept the hypothesis H. We have seen that the Bayesian approach to this is very simple: We report the posterior probability that H is true. This probability is meaningless in the frequency framework, and the frequentist approach to hypothesis testing is more convoluted. First, it is necessary to choose a *rule* for testing, which determines for any given data x whether to accept or reject H. Next, the behavior of this rule must be evaluated in repeated sampling, to find out the probability α that H would be rejected if it is actually true (the probability of "first kind of error"). Finally, if x does indeed lead to rejection of H then we report that H is "rejected at the $100\alpha\%$ level of significance." Otherwise it is "not rejected at the $100\alpha\%$ level of significance."[18]

Notice that the extra complexity of the frequentist approach is necessary because we can only talk of probabilities when they are associated with aleatory uncertainty. Hence, it is necessary to consider all frequentist inferences as instances of inference rules, whose properties are determined by imagining them to be applied in repeated sampling.

One problem with the frequentist formulation is that it is rarely fully understood (even by many practising statisticians, let alone by their clients). When told that H is rejected at the 5% level, this is almost universally interpreted as saying that there is only a 5% chance that H is true. Of course, this cannot be the correct interpretation because it makes a probability statement about the hypothesis (and hence about θ). Only a Bayesian analysis can make such a statement. Yet frequentist inferences are invariably misinterpreted in this way because they seem to make a much simpler and more useful statement ("the probability that H is true is 0.05") than they really do ("if H were true, then the probability that the data would fall in the prespecified region in which they have been observed to fall on this occasion is 0.05").

Similarly, a frequentist confidence interval is nearly always interpreted as a Bayesian credible interval. Thus, the statement that $[1.2, 4.7]$ is a 95% confidence interval for some parameter θ is almost invariably understood as saying that there is a 95% chance that θ lies between 1.2 and 4.7. This cannot be correct because it is a probability statement about θ. The correct interpretation is: that a rule of inference has been applied which yields an interval estimate for θ, that in repeated sampling the intervals constructed by

[16]The distinction between aleatory and epistemic uncertainty is not always clearly delineated in practice. It is even arguable that at a fundamental level true randomness does not exist. Nevertheless, the distinction is useful in discussing the difference between Bayesian and frequentist approaches.

[17]In fact, Your uncertainty about the data x is both aleatory and epistemic. Since the parameters of the data-generating process have epistemic uncertainty, the uncertainty in x is more than just the aleatory uncertainty induced by randomness.

[18]There are actually two different versions of the frequentist hypothesis test. This is the Neyman–Pearson form of significance test. The Fisherian p-value requires a nested set of rejection regions to be defined, and then p is the α value of the region for which the observed data x lie on the boundary.

this rule contain θ with probability 0.95, and that when applied to the particular data x this rule has produced the interval $[1.2, 4.7]$. The confidence interval is generally interpreted as a credible interval because the Bayesian statement is simpler and more natural.

The Bayesian methods answer inference questions in direct and simple ways. The frequentist inferences have more indirect and easily misunderstood interpretations.

Paradoxes in Frequentist Inference

There is much abstract and theoretical debate about the merits of Bayesian versus frequentist methods. In general, the Bayesian approach is seen to be more philosophically consistent, whereas the frequentist approach gives rise to quite paradoxical properties. Rather than dwell in detail on these, we present here just two related instances where frequentist and Bayesian methods behave quite differently, and try to present both sides of the argument in each case.

The Likelihood Principle

To illustrate the difference between Bayesian and frequentist methods, consider again the example of mobile phone usage while driving. We supposed that the data comprised a survey of 100 people, in which 23 admitted to using the phone while driving. The usual frequentist estimate of θ in this situation is $\hat{\theta} = 23/100 = 0.23$. However, this presupposes that the survey size $n = 100$ was fixed and the observation is $r = 23$. If we took repeated samples of 100 people and calculated $\hat{\theta} = r/100$ every time, then on average these estimates would equal the true value of θ; this is the frequentist estimation property known as *unbiasedness*. Suppose, however, that the survey was conducted differently, so that we kept sampling until we obtained 23 people who claimed to have used a mobile phone while driving in the last year. Now $r = 23$ is fixed, and it is $n = 100$ that is random. If we repeatedly took samples, in each case sampling until $r = 23$, and calculated $\hat{\theta} = 23/n$ in each case, then the values we got would not average to θ. In this different kind of sampling, the appropriate frequentist unbiased estimator is $\theta^* = 22/(n-1)$, which in the particular case that we observed of $n = 100$ yields $\theta^* = 22/99 = 0.2222$.

A Bayesian analysis of this problem would be quite different. The posterior distribution would be the same in both cases, so You would obtain the same Bayesian inferences, including estimates, no matter which sampling method was used.

Both frequentists and Bayesians regard this example as favoring their approach. Frequentists assert that if the data are obtained from different sampling methods, then it is obvious that they have different meaning and we should make different inferences. The Bayesian argument is that in both cases we have observed 23 people out of 100 who have used their mobile phones while driving, and knowing whether we fixed 23 or fixed 100 is irrelevant because this knowledge in itself obviously conveys no information about θ.

To add a further twist to this example, suppose that the experiment were conducted in yet another way, with the survey being continued until we ran out of time, money, or energy to continue.[19] Now neither r nor n is fixed. The frequentist theory can have enormous trouble with such a situation, because it may be almost impossible to determine what repeated samples, conducted under the same conditions, would look like. The Bayesian theory has no such difficulty. It is obvious that the mechanism for determining the sample size is not itself informative about θ, and the inference is again the same as if n were fixed at 100, or r at 23.

Formally, Bayesian inference adheres to the Likelihood Principle, which in simple terms says that inference should depend on the data but not on what data we might have obtained; see Berger and Wolpert (1988) for a much more detailed explanation. Different sampling mechanisms lead to different alternative samples. For instance, $r = 23, n = 101$ is possible if r is fixed (or if neither is fixed) but not if n is fixed at 100. Because frequentist methods are evaluated in terms of repeated sampling, they do depend on the

[19] Many actual surveys are in reality conducted like this, even though the scientists may subsequently report them as if the sample size was predetermined!

sampling mechanism. To adherents of the frequentist philosophy, this is natural and unexceptional. To Bayesians, the frequentist approach is illogical in giving inferences that depend on features of the experiment (such as whether $n = 100$ or $r = 23$ was predetermined) that do not in themselves convey any information about the unknown parameters.

Applying Inferences to Particular Data

In a similar way, frequentist inference depends on the rule of inference having been prespecified. Suppose in the mobile phone use example (with fixed $n = 100$), we consider the estimation rule $\theta^+ = r/100$ if r is an odd number and $\theta^+ = r/101$ if r is an even number. Now this rule is not unbiased, and indeed is biased downwards (tending to give estimates that are too low). Given our actual observation, both rules give the same estimate, $23/100 = 0.23$. In one case, however, the estimate results from applying an unbiased estimation rule, while in the other it comes from a biased rule. So is the actual estimate, 0.23, biased or unbiased? Unbiasedness or biasedness is a property of the rule, and in frequentist terms it does not make sense to ask whether an estimate obtained from a particular set of data is unbiased.

Perhaps a more convincing example can be given in the case of a confidence interval. If we are told that $[1.2, 4.7]$ is a 95% confidence interval for θ, then we know that on 95% of the occasions that this rule is used the calculated interval will contain θ. It is now particularly compelling to say that we should give a probability of 0.95 to the interval containing θ on *this* occasion. Admittedly, this is a Bayesian statement, but what is wrong with this very natural transfer of the 95% property from the rule to the instance? The answer is that $[1.2, 4.7]$ could easily also be obtained by applying some other interval estimation rule that is, for instance, a 90% confidence interval. So is the probability 95% or 90% that $[1.2, 4.7]$ contains θ?

From the frequentist perspective, the answer depends again on what might have been obtained but was not, since different rules that give the same inferences on the actual data x would give different inferences on other data. Bayesian inferences apply unambiguously to the particular data that have been observed.

Subjectivity and Science

The most persistent criticism that is made of Bayesian inference is that it is subjective. This is undeniably true, since Bayesian methods are based on the subjective formulation of probability; the posterior distribution represents the beliefs of a particular person (You) about θ. To people who were trained to think that statistical analysis of data must be based on scientific principles, and that science is above all objective, this seems to provide a compelling reason to reject the Bayesian approach. However, closer examination shows—in the author's opinion, at least—that the criticism is vacuous because neither frequentist methods nor science itself is objective.

It is certainly true that science aspires to be objective, and avoids subjective judgments wherever possible. But in every field of science we find controversy and differences of opinion over topics of current interest. The progress of science is achieved through debate, the accumulation of evidence and convergence upon explanations and interpretations of the evidence. Questions that may seem to be resolved in this way can be reopened when new data throw an accepted theory into doubt, or when somebody interposes a new explanation or interpretation, as witness the revolution in thinking that came at the start of the twentieth century with relativity theory superseding the previously accepted Newtonian physics. Consider any piece of research that is published in some eminent scientific journal. The authors will present their data and the conclusions that they draw from those data. The data themselves may be considered to be objective, but the conclusions are not. The authors will describe the process by which the data were collected and describe their own interpretation of those data as clearly as they can, to convince the reader to accept their conclusions. The conclusions may indeed be deemed uncontroversial, but often their fellow scientists will apply their own interpretations (and perhaps reach different conclusions) or else reserve

judgment until the issues have been debated more or until more data are available. Objectivity in science is really a convergence of subjective opinion, and that agreement may be only temporary.[20]

Subjectivity in frequentist statistics is equally easy to see. In practice, two statisticians faced with the same data will often reach different inferences. This may be because they have chosen to use different inference rules, for instance, two different hypothesis tests. The choice of an estimator, a test, or a confidence interval is one source of subjectivity, and although in some simple problems there are universally agreed "best" inferences this is rarely the case in more complex analyses. A more common reason for reaching different conclusions is that the statisticians model the data differently. The effect of this is that they obtain different likelihoods. From the Bayesian perspective, this is entirely natural because all probabilities are subjective and the likelihood is no exception. But from the perspective of a frequentist who criticizes the Bayesian statistician for being subjective, this is an embarrassment. It is this author's contention that no methods of statistics are objective, just as science is not objective.

This is not to say that we should make a virtue of subjectivity. Like science itself, a Bayesian analysis aspires toward objectivity and attempts to avoid those aspects of subjectivity that have given a derogatory connotation to the word "subjective." Thus, probabilities may be subjective, but they should not be affected by prejudice, superstition, or wishful thinking.

Furthermore, Bayesian analysis reflects the above view of the process of science perfectly. It was explained in the section Prior Distribution how the prior distribution has less influence if the data are strong. Thus, as more data are collected, people who might have begun with very different prior beliefs will find that their posterior distributions converge. Eventually, differences of prior opinion are overwhelmed by the accumulating evidence, which is precisely the way that science progresses. Indeed, the fact that Bayesian methods recognize prior opinion is a positive benefit, because it allows us to see when this convergence has taken place. If the data are not strong enough to yield uncontroversial inferences, then this is an important fact that is not apparent in any frequentist analysis.[21] Howson and Urbach (1993) present a detailed argument in favor of Bayesian statistics from the perspective of the philosophy of science.

Bayesian Statistics Today

Modern Bayesian statistics is a rich and powerful framework in which to make inferences and decisions. We consider here a few of the more striking features of Bayesian statistics today.

The Growth of Bayesian Applications

Since about 1990, there has been a dramatic growth in the use of Bayesian methods. In some application areas today, a Bayesian approach is almost a hallmark of leading-edge research. These are often fields where data are scarce, or have complex structures that are difficult to analyze, whereas frequentist methods are still dominant in the more traditional application areas of statistics. In the social sciences particularly, there is more recognition of the role of judgment in interpreting data, and there is less resistance to the apparent subjectivity of Bayesian methods.

For example, the relatively new field of health economics is concerned with assessing the cost-effectiveness of competing medical technologies (such as alternative drugs, surgical interventions, or vaccinations). Such assessments are typically made by assembling evidence on the effects (both positive and negative) of the treatments and the costs incurred (for the treatments themselves and

[20] I have been told by senior scientists that personal judgment does not play a role in their work, but they are wrong. What makes these people leaders of their own fields is that their opinions and judgments are esteemed by their fellows.

[21] Notice, however, that the convergence of opinion relies on all participants agreeing on the likelihood. This mirrors the need for scientists to agree on the interpretation of data before they can agree on the conclusions that can be drawn from them.

any other medical resources used). The resulting evidence base is complex, and inevitably very weak in some areas. Bayesian methods are acknowledged as essential to produce meaningful statistical analyses in such problems. In contrast, frequentist methods are still the dominant methodology in the more well-established field of analyzing clinical trial data. Clinical trials have traditionally produced high-quality, well-structured data, and have been large enough to ensure that prior information and preexisting opinions would be overwhelmed by the trial evidence. Here too, however, Bayesian methods are beginning to become more attractive, partly driven by the high cost of modern drug development that has led to a desire for smaller trials and more efficient use of all available information.

In the following subsections, we look at some of the factors that have played a part in stimulating this rapid growth in the uptake of Bayesian methodology.

Bayesian Computation

Two distinct steps can be identified in the basic Bayesian method:

1. *Bayesian Modeling*. Identify the unknown parameters and the inference questions about these parameters that are to be answered. Construct the likelihood and the prior distribution to represent the available data and prior information.

2. *Bayesian Analysis*. Obtain the posterior distribution and derive inferences.

We will consider Bayesian modeling in the section One Coherent Framework for Thinking, so concentrate here on the second step, Bayesian analysis.

Until the advent of powerful computational tools, Step 2 represented a major difficulty except in very simple problems. To illustrate these difficulties, first suppose we have a sample of data from a normal distribution with unknown mean μ and known variance σ^2. The unknown parameter that we have generically denoted by θ is, in this example, μ. The likelihood for this sample can be written[22]

$$f(x\,|\,\mu) \propto \exp\left\{-\tfrac{n}{2\sigma^2}(\mu-\bar{x})^2\right\}, \qquad (6.2)$$

where $\bar{x}$ is the sample mean. Now suppose that the prior distribution for μ is normal with mean m and variance v, so that

$$f(\mu) \propto \exp\left\{-\tfrac{1}{2v}(\mu-m)^2\right\}. \qquad (6.3)$$

This is an instance of what is called a conjugate prior distribution, because it combines nicely with the likelihood to produce a posterior distribution that is very easy to analyze. In fact, the combination of likelihood (6.2) and prior (6.3) is easily shown to result in a posterior distribution for μ that is also normal. Derivation of inferences such as the posterior mean or credible intervals is now simple. In this case, the whole Bayesian analysis can be done analytically because the posterior is found to have a well studied, standard distributional form.

However, if the prior distribution is not normal, the posterior will typically no longer be so simple. For instance, if the prior density has the logistic form

$$f(\mu) \propto \exp(p\mu)\,\{1+\exp(\mu)\}^{-(p+q)},$$

then the posterior will not have any standard form. To derive any inference such as the posterior mean or a credible interval will now require numerical computation. Because there is only one parameter in this problem, these calculations require only numerical integration in one dimension, which is straightforward. In the period from the birth of modern Bayesian thinking in the 1950s to at least the mid-1980s, Bayesian analysis was restricted to situations in which conjugate prior distributions were available, or where the number of parameters was small enough for computation of posterior inferences by numerical integration to be feasible.

[22] We have simplified the likelihood here by writing it as *proportional to* the expression shown. That is, factors that do not depend on the parameter μ have been dropped. We do the same with the prior distribution in the next expression. These are legitimate simplifications because Bayes's theorem says the posterior distribution is proportional to the product of prior and likelihood, and this remains true after removing any such constant factors in either term.

Problems that could be analyzed routinely by frequentist methods, such as generalized linear models with many explanatory variables, were outside the reach of Bayesian methods.

This changed with the development of the computational technique known as Markov chain Monte Carlo, universally abbreviated to MCMC, so that we can now perform those computations even in very complex, multiparameter situations.

MCMC is based on two conceptually very simple ideas. The first is that of sampling-based computation. Suppose that we wish to compute the posterior mean of the parameter θ_1, which is the first element of the vector θ of, say, k parameters. Formally, this is

$$E(\theta_1 \,|\, x) = \int \theta_1 f(\theta \,|\, x)\, d\theta$$

and involves integrating over the whole k-dimensional space of the parameter vector θ.[23] If k is more than about 10, this is a very substantial computation using numerical integration. However, imagine that we could take a sample of N values from the posterior distribution $f(\theta \,|\, x)$. Denote these by $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(N)}$. Then we would in particular have a sample of values of the first parameter θ_1, obtained by taking the first element in each of the vectors $\theta^{(i)}$, $i = 1, 2, \ldots, N$. We could use the sample mean $\bar{\theta}_1$ as an approximation to $E(\theta_1 \,|\, x)$. If the sample were very large, for instance $N = 10^6$, then we could regard this as effectively an accurate *computation* of $E(\theta_1 \,|\, x)$.

Direct sampling like this from the posterior is sometimes feasible, even in some quite large and complex problems, and is referred to as Monte Carlo computation. However, in most serious applications of Bayesian analysis the posterior distribution is too complex and high dimensional for this direct approach to be feasible. We then employ the second device, which is based on the theory of Markov chains. We again obtain a series of vectors $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(N)}$, but these are not sampled directly from $f(\theta \,|\, x)$ and they are not independent. Instead, each $\theta^{(i)}$ depends on the previous $\theta^{(i-1)}$ and is sampled from a distribution $g(\theta^{(i)} \,|\, \theta^{(i-1)})$. This means that the $\theta^{(i)}$s are a Markov chain. The conditional distribution g, which is known as the *transition kernel* of the chain, is chosen so that for sufficiently large i the distribution of $\theta^{(i)}$ converges to the posterior distribution $f(\theta \,|\, x)$.[24] Markov chain theory provides relatively simple criteria under which this convergence will occur, and in practice there are numerous ways of constructing a suitable transition kernel to sample from any desired posterior distribution.[25]

The combination of the two ideas of sample-based computation and the Markov chain is MCMC. To go more deeply into the technique of MCMC would require more than this chapter, and indeed whole books have been written about it (see, for instance, Gilks, Richardson, & Spiegelhalter, 1995). Instead, we will just note the following important points:

- Although convergence is guaranteed eventually, it is not possible to say how large a sample must be taken before successive values can be considered to be sampled from the posterior distribution. Judging when the sample is large enough is something of an art, although there is a growing body of diagnostics to help with this task.
- Successive points in the Markov chain are correlated, and the strength of this correlation is very important. A highly correlated chain converges slowly and moves around the parameter space slowly, so that a larger

[23] We often think of this integration in two stages. First, we integrate $f(\theta \,|\, x)$ over all elements of θ except θ_1, a $(k-1)$-dimensional integration, to obtain the marginal density $f(\theta_1 \,|\, x)$. Then we integrate with respect to θ_1 to obtain its posterior mean as $E(\theta_1 \,|\, x) = \int \theta_1 f(\theta_1 \,|\, x)\, d\theta_1$.

[24] This is true no matter what the initial value $\theta^{(1)}$ is. In simple terms, the chain can be said to have converged when it has "forgotten" where it started from.

[25] To understand MCMC, it is helpful to think first of simple Monte Carlo sampling as, for example, shooting randomly into the space of possible values of θ. Each shot is distributed according to the posterior distribution and successive shots are independent. In contrast, MCMC starts at an arbitrary point $\theta^{(1)}$, and then wanders around the space, each successive value being a random move away from the previous one. If the transition kernel is appropriately chosen, this wandering point will make its way into the part of the θ space with appreciable posterior density, and will spend more time in regions with higher density and less in regions with lower density, so that the collection of points behaves like a sample from the posterior distribution.

sample is needed to compute relevant inferences accurately. Devising a chain that has relatively low correlation is another task that is something of an art.

The ability of MCMC to tackle extremely complicated problems with very large number of parameters is a major factor in the growth of applied Bayesian statistics. As remarked above, the practice of MCMC is still under rapid development and is a skilled task. There is a powerful software package available, known as WinBUGS,[26] but this also requires a relatively sound knowledge of MCMC practice. As yet it is not truly easy to use software for Bayesian computation.

One Coherent Framework for Thinking

Another appealing feature of the Bayesian approach is its conceptual simplicity and consistency. In effect, a Bayesian analysis involves only the formulation and manipulation of probabilities. The process of building a Bayesian model is all about formulating beliefs in terms of probabilities, and it does not matter whether these probabilities represent aleatory or epistemic uncertainties. The second step of the Bayesian method, which is the derivation of the posterior distribution and inferences, is in principle simply a matter of manipulating probability distributions. The key requirement for an applied Bayesian statistician is to be able to think in terms of probabilities formulating knowledge and uncertainties.

The frequentist philosophy is different. Probabilistic modeling is used to create the likelihood, but the formulation and choice of inference rules are based on an array of more or less ad hoc criteria for what constitutes a good rule. The frequentist statistician is free to propose new rules, and unless they are demonstrably and uniformly inferior to another rule (which can rarely be shown) it is legitimate to use them.

This aspect of Bayesian inference as a coherent framework for thinking about uncertainty emerges very clearly in the following example, where there are no aleatory uncertainties at all. Mathematical models are widely used in science, engineering, economics, and other fields to describe real-world processes, with a view to understanding and predicting their behavior. Such models are usually implemented in computer programs, which can be very large and take anything from a few seconds to many hours to run. In practice, the user of such a model does not require to run it just once, but wishes to consider what outputs are predicted by the model for a variety of settings of its inputs. In some cases, the number of runs that would, in principle, be required is so large that it is impractical to do so within any realistic time-span. Bayesian methods have been developed to enable such analyses to be done without physically running the model for all the necessary input combinations.[27] The idea is to model uncertainty about what outputs the model would produce at some input settings for which the model has not actually been run. There is no aleatory uncertainty because the model itself is deterministic; running it at any given inputs will always produce the same outputs. However, there is clearly epistemic uncertainty about what the outputs will be before we actually run the model. Bayesian methodology can model this uncertainty as if the relationship between inputs and outputs was a random function.

Design, Decision, and Prediction

We end this chapter by highlighting several kinds of problems where a Bayesian approach is more natural and powerful than frequentist methods. The design of experiments and observational studies is obviously such an area, because before we actually collect the data there is only prior information. Frequentist methods must use prior information, but do so in an informal and oblique way. In a Bayesian approach, the prior information is explicit and is used to identify optimal designs.

Decision theory is a large topic. The most difficult decisions arise when there is uncertainty

[26] See http://www.mrc-bsu.cam.ac.uk/bugs.

[27] See O'Hagan (2006).

about the consequences of our actions. The uncertainty about consequences is invariably (at least partially) epistemic and cannot be addressed by frequentist methods. A Bayesian approach is quite natural in this situation, and decision makers rarely exhibit any resistance to the idea that such uncertainties should, in principle, be expressed as probabilities. Bayesian decision theory chooses the optimal decision by maximizing the expectation of a utility function that represents the value of different consequences for each possible decision. The expectation in question is taken with respect to the uncertainty in the consequences. The Bayesian development of optimal experimental designs is actually an instance of Bayesian decision theory.

Finally, consider the prediction of future data. Suppose that data on the efficacy of some medical treatment in a sample of patients have been obtained in a clinical trial, and a clinician wishes to predict the response of new patients to this treatment. This is another area where frequentist methods have difficulty. The uncertainty in future data is primarily epistemic. Frequentist approaches introduce aleatory uncertainty by regarding the new patient as randomly chosen from the population of all potential patients, but there is still epistemic uncertainty because, despite the clinical trial data, there is still uncertainty about the true mean efficacy of the treatment.

Conclusion

This chapter has tried to explain the essence of the Bayesian approach to statistics, how it differs from the frequentist approach and what advantages have caused it to grow dramatically in usage since the late 1980s. The presentation has not been completely impartial because the author has been firmly committed to the Bayesian framework for more than 30 years. It is also important to recognize that within the community of users and advocates of Bayesian methods there is a diversity of opinion on some issues that could not be fully covered within this chapter. The reader is advised to seek other opinions, to which end there are recommendations for further reading at the end of the chapter.

References and Further Reading

Berger, J. O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, *1*, 385–402. (Also available from http://ba.stat.cmu.edu/vol01is03.php)

Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle*. Haywood, CA: The Institute of Mathematical Statistics. (An authoritative text on the subject, although now a bit old)

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory*. New York: Wiley. (A deep and authoritative text with a huge bibliography. Takes a different stance from that presented here on a number of issues)

Berry, D. A. (1996). *Statistics: A Bayesian perspective*. London: Duxbury. (Teaches a substantial amount of Bayesian statistical methods using only elementary mathematics, with an emphasis on medical applications)

Congdon, P. (2001). *Bayesian statistical modelling*. Chichester, UK: Wiley.

Congdon, P. (2003). *Applied Bayesian models*. Chichester, UK: Wiley. (This and the preceding book concentrate on Bayesian modeling and computation in real, sometimes complex situations)

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1995). *Markov chain Monte Carlo in practice*. London: Chapman & Hall. (Although not up-to-date with many modern developments, this is an excellent introduction to the techniques of Markov chain Monte Carlo)

Goldstein, M. (2006). Subjective Bayesian analysis: Principles and practice. *Bayesian Analysis*, *1*, 403–420. (Also available from http://ba.stat.cmu.edu/vol01is03.php)

Howson, C., & Urbach, P. (1993). *Scientific reasoning: The Bayesian approach* (2nd ed.). Chicago: Open Court. (A book that goes deeply into philosophy of science, but generally quite readable)

Lee, P. M. (2004). *Bayesian statistics: An introduction* (3rd ed.). London: Edward Arnold. (A basic text using college level mathematics)

Lindley, D. V. (1980). *Making decisions* (2nd ed.). New York: Wiley. (Deals with subjective probability and decision making in a very clear and non-technical way)

Migon, H. S., & Gamerman, D. (1999). *Statistical inference: An integrated approach*. London: Edward Arnold. (A modern and concise text handling both Bayesian and frequentist theories at an intermediate level)

O'Hagan, A. (1988). *Probability: Methods and measurement*. London: Chapman & Hall. (An elementary but rigorous treatment of subjective probability, leading into exchangeability and basic statistical ideas)

O'Hagan, A. (2006). Bayesian analysis of computer code outputs: A tutorial. *Reliability Engineering and System Safety*, *91*, 1290–1300. (Also available from http://mucm.group.shef.ac.uk)

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., et al. (2006). *Uncertain judgements: Eliciting expert probabilities*. Chichester, UK: Wiley. (A survey of the elicitation of subjective probabilities to build prior distributions)

O'Hagan, A., & Forster, J. J. (2004). *Bayesian inference* (2nd ed., Vol. 2B). London: Edward Arnold. (Assumes a strong background in mathematics and statistics but in a readable style)